Classificando assuntos no dataset de artigos da BBC

Henrique Rocha Bomfim

I. Section 1: Dataset

ESTE estudo examina um banco de dados de artigos da BBC classificados em cinco categorias e avalia a performance de um classificador, visando melhorias. O dataset, descrito em [1], contém 2.126 instâncias distribuídas em cinco categorias rotuladas, com as colunas ”category” e ”text” armazenando as informações.

O dataset foi obtido no Kaggle através de uma busca por ”text categorization” e filtragem de Notebooks e Datasets, sendo então utilizado para as análises deste estudo.

II. Section 2: Classification Pipeline

Nesta seção, o pipeline de classificação foi configurado para pré-processar o texto e treinar modelos de machine learning. O texto foi limpo, removendo-se stopwords e aplicando lematização com a biblioteca WordNet. As palavras foram vetorizadas usando Bag-of-Words (BoW), que considera a presença de termos específicos importante para a categorização, algo necessário para este problema de classificação pois certas palavras representam mais um assunto do que outros, o que ajuda a melhorar a performance no momento de decidir o assunto do artigo.

Os modelos utilizados foram Regressão Logística e Naive Bayes, ambos treinados para distinguir entre as categorias. Embora BoW seja eficaz para identificar palavras-chave de cada categoria, essa técnica pode ser suscetível a confusões, como no caso de palavras ambíguas, como ”game”, que pode se referir tanto a esportes quanto a entretenimento.

III. Section 3: Evaluation

Os modelos foram avaliados com métricas como acurácia balanceada, devido ao possível desbalanceamento do dataset. As divisões de treino e teste foram feitas aleatoriamente, com estratificação para manter a representação das categorias. Múltiplas iterações foram realizadas, e o Naive Bayes teve melhor desempenho em categorias mais distintas, como esporte e tecnologia, enquanto houve mais confusão entre negócios e política. Também tendo ligeira vantagem em precisão sobre a Regressão Logística. Além disso, foi classificada uma coincidência em game, mr e year, conforme Figura 1. As palavras parecem representar bem 5 possíveis tópicos, levando em consideração as listas de palavras presentes na Figura 1.

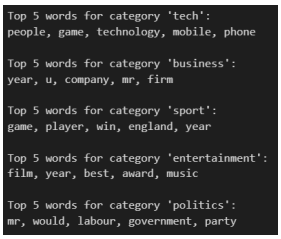

Fig. 1. Listagem das principais palavras para cada categoria identificada.

IV. Seção 4: Tamanho do Dataset

Tanto a Regressão Logística quanto o Naive Bayes apresentaram erros menores à medida que o tamanho do dataset aumentava, indicando que mais dados melhoram a performance. No entanto, aumentar o número de exemplos não necessariamente melhora a acurácia para todas as categorias devido à sobreposição de tópicos. Expandir o dataset é desafiador, pois há um limite para a quantidade de artigos da BBC disponíveis anualmente. O gráfico (Figura 2) ilustra como as taxas de erro mudam com tamanhos de dataset variados. Adicionar alguns milhares de instâncias seria vantajoso nesse primeiro momento, especialmente porque algumas categorias da análise de tópicos ficaram sem dados suficientes, conforme será discutido na seção 5. Quando a quantidade de dados for grande o bastante para os erros de treino e teste terem valores muito próximos na primeira casa decimal, aumentar os dados não será mais relevante.

Fig. 2. Gráfico comparativo da taxa de erro conforme aumento do uso do dataset.

V. Section 5: Topic analysis

Ao solicitar que o classificador de duas camadas encontrasse 5 tópicos, apenas 2 tópicos foram encontrados e testados, com cerca de 97% de acurácia, contra 60% para business e 81% para entretenimento da Classificação. Os demais 3 tópicos gerados não encontraram resultados próximos o suficiente na base de dados, então não foram considerados. Esse resultado poderia ser melhor com mais dados ou com um menor número de tópicos esperados. Ou seja, o sistema de classificação foi melhor nos demais tópicos.

REFERENCES

[1] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 377–384, ISBN: 1595933832. DOI: 10.1145/1143844.1143892. [Online]. Available: https://doi.org/10.1145/1143844.1143892.